From Goals to Actions: Designing Context-aware LLM Chatbots for New Year's Resolutions

Yan Xu
Meta
Reality Labs Research
Redmond, Washington, USA
yanx@meta.com

Brennan Jones
University of Toronto
Faculty of Information
Toronto, Ontario, Canada
brennan.jones@utoronto.ca

Hannah Nguyen
Meta
Reality Labs Research
Redmond, Washington, USA
hannahnguyen@meta.com

Qisheng Li
Meta
Reality Labs Research
Redmond, Washington, USA
qishengli@meta.com

Stefan Scherer
Meta
Reality Labs Research
Redmond, Washington, USA
stefanscherer@meta.com

Abstract

When pursuing new goals, people often struggle to determine what actions to take. Large-language-model (LLM) chatbots can provide information and interactivity, and combining them with context awareness could enhance the relevance and proactivity of action recommendations. However, there is a gap in understanding the role that such technologies can play in taking a holistic view of the user's multiple goals, complex contexts, and constraints over time. We developed a technology probe of a personalized context-aware LLM chatbot and deployed it with 14 participants for 2-4 weeks for their 2024 New Year's resolutions. We observed users achieve a high adoption rate of actions and greater success in the pursuit of goals in the first week, as well as the rapidly evolving user needs over time. We discuss how to best leverage context-awareness for AI agent design, and the novel roles that AI could adopt for an ecosystem of services and agents.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; **Natural language interfaces**.

Keywords

chatbots, agents, language models, goal pursuit, behavior change, context-aware computing

ACM Reference Format:
Yan Xu, Brennan Jones, Hannah Nguyen, Qisheng Li, and Stefan Scherer. 2025. From Goals to Actions: Designing Context-aware LLM Chatbots for New Year's Resolutions. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces (CUI '25), July 08–10, 2025, Waterloo, ON, Canada.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3719160.3736637

CUI '25, Waterloo, ON, Canada

ACM ISBN 979-8-4007-1527-3/25/07
https://doi.org/10.1145/3719160.3736637

1 Introduction

Every year 20%-30% adults all over the world make New Year's resolutions (NYRs) such as well-being, financial and learning goals [2, 42]. Among people who have enough motivation to make NYRs, many fail by February (20% according to [38], 45% according to [37]). Lack of specificity is one of the top reasons [38]. For example, a goal like *"I want to be healthier"* does not translate to what to do directly. As proven by Goal-Setting Theory and its related studies [31–33], planning specific, realistic, and achievable actions is crucial towards larger goals [23, 30]. To figure out what to do for their new goals, people often turn to sources such as the Internet, friends, or family to seek advice. But others may not be able to tailor suggestions to one's own circumstances because they are not aware of the specific daily circumstances that the individual experiences, let alone the dynamically changing contexts and constraints that may affect one's desire and ability to achieve their goals (which can include the individual's other goals or daily responsibilities).

In recent years, people started to seek advice from large language models (LLMs) [45]. Research shows a perceived higher quality of advice compared to traditional advice columns [15]. LLMs can achieve generalized knowledge about a variety of topics and can also be fine-tuned or tailored to acquire specialized knowledge about a particular domain or goal [28]. However, there are limitations to seeking advice from LLMs. First, the quality of the advice depends on the quality of the prompts that people create. However, many users are not experts in crafting detailed and effective prompts. In a user study with graduate students, researchers, and designers [46], it was found that users struggle to generate prompts, evaluate prompt effectiveness, and explain prompt effects. Second, standard LLM chatbots are not adaptive to the user's individual attributes or real-time contexts, and are merely reactive to the user's prompts rather than proactively guiding the user at opportune moments. Users may not always be aware of when they are in a good opportunity to act or discover a new action toward their goals.

Context-aware computing can understand different aspects about a user's circumstances, and use this understanding to recommend contextually-relevant actions that the user can take toward their goals. Technologies such as smartphones, smart glasses, and smartwatches can capture context, including *environmental context* (e.g., from one's immediate environmental surroundings, captured via

sensors or fetched from the Internet) and *personal context* (e.g., related to one's goals, interests, availability, capability, personality, preferences, and behaviors), and use this to generate and deliver action recommendations at the right moments, thus providing a *just-in-time* approach [36, 44]. Context-aware recommendations combined with LLMs have the potential to provide users with personalized guidance that considers a holistic view of users' complex contexts and all of their goals, especially for new goals and unfamiliar contexts. However, while there is a plethora of work on context-aware technological interventions for behavior change, most current approaches focus on single goals (e.g., *"walk 10,000 steps per day"*) or single goal categories (e.g., health, fitness). Additionally, most context-aware systems focus only on limited context variables such as location or step count. There is a gap in understanding the role that such technologies could play in taking a *holistic view* of the user's contexts and constraints over time, as well as the complex relationship between the user's multiple goals and their evolving goal journeys.

To achieve this understanding, we conducted a field study where we deployed an LLM-based chatbot app as a technology probe [16] supporting 14 participants' genuine 2024 NYRs for two to four weeks. The chatbot app proactively recommends actions to the user for their goals, based on context factors that can be detected or inferred by the user's phone (e.g., location, time, and weather), the user's personal profile (e.g., goals, preferences, tools, etc.), and the communication history users have with the chatbot. Users can also converse with the chatbot freely, to ask for more details or clarification on a recommended action, ask for new recommendations, edit their contexts, change goals, set reminders, etc.

Our longitudinal field deployment enables us to dive into the harder part of long-term behavior change: whether and how people stick with their goal pursuit and how effective AI chatbots are in the long run. Our longitudinal data showed a trend where contextualized action recommendations were most helpful in the first week. The technology probe helped people kickstart the pursuit of their goals with relevant, feasible and personalized actions that users may not have thought about before, without the pressure of social comparison. However, its effectiveness reduced overtime, revealing the evolving user needs as people start to habitualize actions. We further analyzed the variance among the users' goal pursuit, which was associated with chatbot engagement, context sharing, and tool integration. Based on the findings, we discuss areas of improvement to better support people's NYR journeys and how to mitigate privacy concerns when contextual information is needed for better assistance. We broaden the design landscape by exploring more holistic roles that AI agents could play.

2 Related Work

2.1 Behavior Change Theories and their Applications to Technological Interventions

Behavior change is a topic that has been long studied in HCI research [18], often with the goal of building *persuasive technologies*, or technologies that designed to invoke habit formation and help the user pursue long-term goals (e.g., [6]). Research on persuasive technologies has suggested that they should be designed to "consider the *practical constraints* of users' lifestyles" [6], which can include individuals' immediate *contexts* and *constraints*, such as their daily schedules, home environments, resources in their homes, personal skills and capabilities given their contexts. This is also supported by the Fogg Behavior Model [11], which suggests that three things are needed for an individual to take action toward a goal: *motivation*, *ability*, and a *prompt*. Another related theory is the COM-B Model [34], which suggests that individuals need *motivation*, *opportunity*, and *capability* to perform a target behavior. Applying behavior science principles like the COM-B model has been shown to enhance the effectiveness of LLMs in activities such as coaching [14], and provides a valuable framework for our system design. Applying both the Fogg Behavior Model and COM-B Model, we designed our chatbot to learn about the user's *motivations* through their profile and their current *abilities* (or *capabilities*) through their present context, so that it can generate *prompts* to the user to make them aware of *opportunities* to perform contextually relevant actions toward their goals in the present moment. Such an approach is similar to other *"just-in-time"* interventions that consider users' contexts when timing and delivering guidance [36]. "Just-in-time" interventions have been employed for goals such as fitness and health management by leveraging factors like location, time of day, physical activity, emotional state, the presence of other people, and medical data [41].

The Transtheoretical Model (TTM) [39] suggests that behavior change occurs in six stages: *precontemplation*, *contemplation*, *preparation*, *action*, *maintenance*, and *termination*. Bak et al.[3] found that LLM chatbots are effective in identifying user motivation states in the later stages of the TTM; but they perform poorly in the early stages. This highlights the need for advanced personalization in LLMs to better serve users at different motivational stages. Our chatbot aims to motivate users particularly in the early and middle stages of the TTM by helping them identify opportunities to take even small actions toward their goals, even when such opportunities are not obvious to them, and by helping users understand why and how such actions could be beneficial to users.

2.2 The Roles of AI and LLMs in Goal Pursuit and Behavior Change

Recent studies have begun to explore leveraging AI, especially LLMs, in the domain of behavior science, which is closely related to our investigation into the use of personalized, context-aware LLMs for supporting goal pursuit. AI conversational agents and chatbots have been employed for achieving goals such as exercising, losing weight [22]), and mental wellbeing [21]. AI agents have also been employed to support journaling and goal reflection [21]. One of the major benefits of AI agents for such purposes is that, as such agents become more advanced and capable of human-like behavior and cognition (especially as they are powered by more advanced models such as LLMs), users can interact more naturally with them through conversation, as they would with another human. This leads to such agents becoming able to play a variety of the same roles that other humans can in one's goal pursuit — e.g., *personal assistant* (e.g., for scheduling or finding recommendations) [25], *coach* or *mentor* [14], *emotional support provider* [29], or *social partner* for progress tracking and casual conversation [24].

Recent studies further illuminate the role of LLMs in motivating physical activity and providing advice. For instance, Jörke et al. [20] develop a LLM-based conversational agent, GPTCoach, and find that while GPTCoach can maintain a supportive coaching tone, its ability to proactively utilize data for motivation varies, emphasizing the importance of personalization and the effective use of data for motivation, aligning with our research on personalized, context-aware LLM chatbots. Wester et al. [45] investigate the influence of user characteristics such as personality and technology readiness when perceiving LLM-generated advice, suggesting that the style of LLM-generated advice can significantly affect user receptiveness. Howe et al. [15] present a compelling comparison between advice from ChatGPT with that from professional columnists, with ChatGPT often viewed as more empathetic and helpful. This supports the potential of LLMs to express empathy when giving users advice, prompting us to further optimize LLM outputs for empathy and relatability in advisory contexts.

Many of these technologies focus on single goals or single goal categories (e.g., health, sustainability). One existing gap is the role that behavior change technological interventions can play for people who have multiple goals or complex contextual circumstances. In other words, *how can technologies for behavior change, including AI agents such as chatbots, take a* ***holistic view*** *of users' daily contexts and the complex relationships between their multiple goals?* This is a key question that we explore through the development of our technology probe and its deployment in our field study. Such an understanding is particularly important, as existing technologies that focus only on single goals are often strong at considering the *motivational* aspect of the Fogg Behavior Model and COM-B Model, but fall short in delivering timely nudges that consider users' complex circumstances, their limited and unique *capabilities* that arise in such circumstances, and the unique *opportunities* that users could discover from their complex circumstances, which includes the interplay of their multiple goals.

3 Technology Probe

Technology probes are simple, flexible, adaptable technologies that are created to understand user needs in real-world settings, field test technologies, and discover design opportunities [16]. These three goals of technology probes fit well with our own research goal, especially considering the emergent and adaptive behaviors that both humans and LLMs may exhibit during their goal pursuit in the real world.

Therefore, we designed and developed a simple yet flexible chatbot that leverages an LLM and contextual information to help users pursue their goals (see Figure 1 for the system information flow and Figure 2 for the chat interface). With this prototype, a user receives contextualized action recommendations throughout their day. They can use the chat interface to further understand, discuss, or keep track of their actions and goals with the agent, which remembers their interaction history and profile. As we discuss in the next section, the choice of the chat interface was a result of iterative testing, which highlighted the importance of flexibility.

3.1 Design Process of the Technology Probe

We developed this prototype by following an iterative design process, working from lower to higher-fidelity prototypes, through four stages of iteration. In the first stage, we produced low-fidelity interface sketches of a mobile action recommendation app with menus and icons, and received and implemented feedback from colleagues in our organization. In the second stage, we implemented a context-aware action-recommendation app that did not have a chat interface, but instead generated and delivered action recommendations to users proactively in the background and reactively whenever the user tapped a button. We demonstrated and tested this iteration with internal users, and learned that users wanted to be able to explain their needs and constraints to the agent. Thus, we determined that, in addition to automatic context awareness, a chat interface, where users could flexibly and explicitly describe their needs, was necessary. We implemented a chat-based interface in the third iteration, demoed it to 13 internal users, and had three additional internal users dogfood it over a period of several weeks for their own goals and daily contexts. From these users, we learned more about what types of queries users wanted to give and what users expected in the agent's outputs. We thus generated the following design guidelines, which we then applied to the design of the latest iteration of our technology probe.

3.2 Design Guidelines for the Technology Probe

(1) **Personalization: Adapt to users' goals and attributes and remember all past interactions with the user.** Like coaches and trainers who keep record of their clients' training histories, the app should remember the interests, goals, capabilities, constraints, and preferences of the user. This information can come from users' explicit sharing or be extracted from previous interactions.
(2) **Contextualization: Adapt to users' real-time and changing contexts.** The app should be observant to multiple context factors that are relevant to their goal pursuit. Proper planning and breaking down of actions into feasible steps can lead to more successful goal pursuit [23], and LLMs are well-poised to tap into users' contexts and personal profile attributes to generate personalized recommendations that are detailed, specific, and feasible for the user in their present context. Lastly, the app should be able to account for inaccuracies and missing information, and any information that the user explicitly gives the app should take precedence over information that is observed automatically by the app.
(3) **Shared agency: Provide assistance both reactively (user-initiated) and proactively (system-initiated).** Users may not always know the best moments to act toward their goals. Thus, having the app continuously observe their contexts and give users insights on opportunities they have to act toward their goals, even in new or non-routine contexts, could provide benefit.
(4) **Flexibility: Accommodate a variety of user queries.** As a technology probe for a field deployment, we cannot anticipate all the usages and user needs. We need a flexible prototype that can capture them. A user may have different

styles, attributes, or level of details when it comes to interacting with the prototype. Thus, we need to capture and respond to a wide range of users' needs and queries, and we decided on a conversational interface (chatbot) as the main UI for this.

Following the development of the latest iteration of our technology probe, we deployed the probe in the field over an extended time period to gather real-world usage, user needs, and feedback from users.

4 System Design and Implementation

4.1 Privacy Protection and Constraints

The implementation of our prototype strictly followed our internal privacy policies and review process, which required the use of internal servers for all AI models, meaning we could only deploy the prototype to devices that had access to the internal server. In our case, these devices were limited to employees' company phones, which also constrained our participants to be internal employees only. Furthermore, we were not permitted to access participants' work calendars for our prototype. Instead, we assumed the work hours for the participants, who all had full-time jobs.

4.2 System Data Flow

Figure 1 illustrates the design of our chatbot app. We designed the system as an 'always-on' app that can be used in the foreground (through the chat interface) and in the background as the user goes about their day. The user first creates a profile by completing a Qualtrics survey, where they provide the following information:

- **Nickname (a name that the system uses to refer to the user)**, **age**, and **gender** (all optional)
- List of the user's **current long-term goals** that they are actively pursuing at the time of the study
- Advantages and disadvantages (i.e., **attributes**) in pursuing long-term goals. Examples could include:
  - *"I am very good at time management."*
  - *"I have children, and so tend to get busy as a parent."*
- List of **tools** that the user has access to at *home*, *work*, and *other places* that are relevant to the user's goals

The user can open the app to ask the chatbot questions at any time. The chatbot continuously tracks the user's context (including their location, time, weather, etc.), and always considers this context and the user's profile when responding to the user and generating recommendations. This context tracking also runs when the app is in the background, so that the chatbot can proactively message the user with recommendations, even when the user is not actively using the app.

The app UI is composed of two primary views. The **chat view** (Figure 2, left) is the space where the user converses back and forth with the chatbot. In this view, action recommendations from the chatbot appear with hyperlinks, which when clicked direct the user to the **action details view** (Figure 2, right). This view provides a space for the user to view 'structured details' about an action recommended to them by the chatbot, including (i) a list of *steps* to complete the action, (ii) the *goals* that the action supports, (iii) a description of which of the user's current *context factors* and *profile attributes* enable the action to be completed and how they enable the action, and (iv) a ranking of the amount of *effort* expected for the user to complete the action.

Using the chat view, the user can converse with the chatbot as they would with a traditional LLM-based chatbot. In addition, the user can perform the following actions directly in the chat interface:

- Ask the chatbot to provide a **new action recommendation**. In this case, the chatbot responds with a new message with a recommended action, based on any constraints that the user provides.
- Specify to the chatbot any **changes in their goals**, including new goals or changes to existing goals. In this case, the chatbot updates the user's goals in their profile.
- **Change their current context** (e.g., tell the chatbot where they are located or what the weather is). In this case, the context that the user specifies would override any context that the system detects or infers.
- **Log an action** as 'completed', 'will consider', 'liked', or 'disliked'. In this case, the chatbot saves the action to the user's personal action log, which they can then view on a separate page in the app at any time.
- Ask the chatbot to **remind them about an action** at a later specified time or when the user is in a specified context (e.g., when the user is at home). In this case, the chatbot flags the action, and when it detects that the specified time arrived or the user is in the specified context, it sends a reminder about the action to the user.

4.3 Implementation

The client app runs on Android and was implemented using React Native. The server was implemented using Node.js, stores the user's profile, and uses GPT-4 Turbo as the chosen LLM, running on a Microsoft Azure cloud instance.

Whenever the user sends a message to the chatbot, the app sends to the server a JSON object containing the message text, the phone's GPS location, the phone's system time, and a flag that indicates if the user is currently located at their 'home' location, 'work' location, or neither. The server then uses the GPS location to fetch the current weather conditions at the user's location using the US National Weather Service (NWS) API[1]. This API returns the current temperature and a short description of the current weather conditions (e.g., *"Sunny"*, *"Overcast with a chance of rain"*, *"Thunderstorms"*). Next, if the user is not currently located at their 'home' or 'work' addresses, the server uses the OpenStreetMap (OSM) Nominatim API [2] to fetch a description of the user's location using reverse geocoding (e.g., *"library"*, *"coffee shop"*, *"gym"*, *"shopping mall"*, etc.). The server also uses the OSM Overpass API[3] to fetch *nearby 'place types'* (e.g., *"cafe"*, *"library"*, *"shopping mall"*) within a 400-meter radius of the user's location.

The server then compiles a prompt and sends to the LLM the message that the user typed along with the user's profile, conversation history (messages older than 24 hours are summarized, recent

[1] US National Weather Service API: https://www.weather.gov/documentation/services-web-api
[2] OpenStreetMap Nominatim API: https://nominatim.org/
[3] OpenStreetMap Overpass API: https://wiki.openstreetmap.org/wiki/Overpass_API

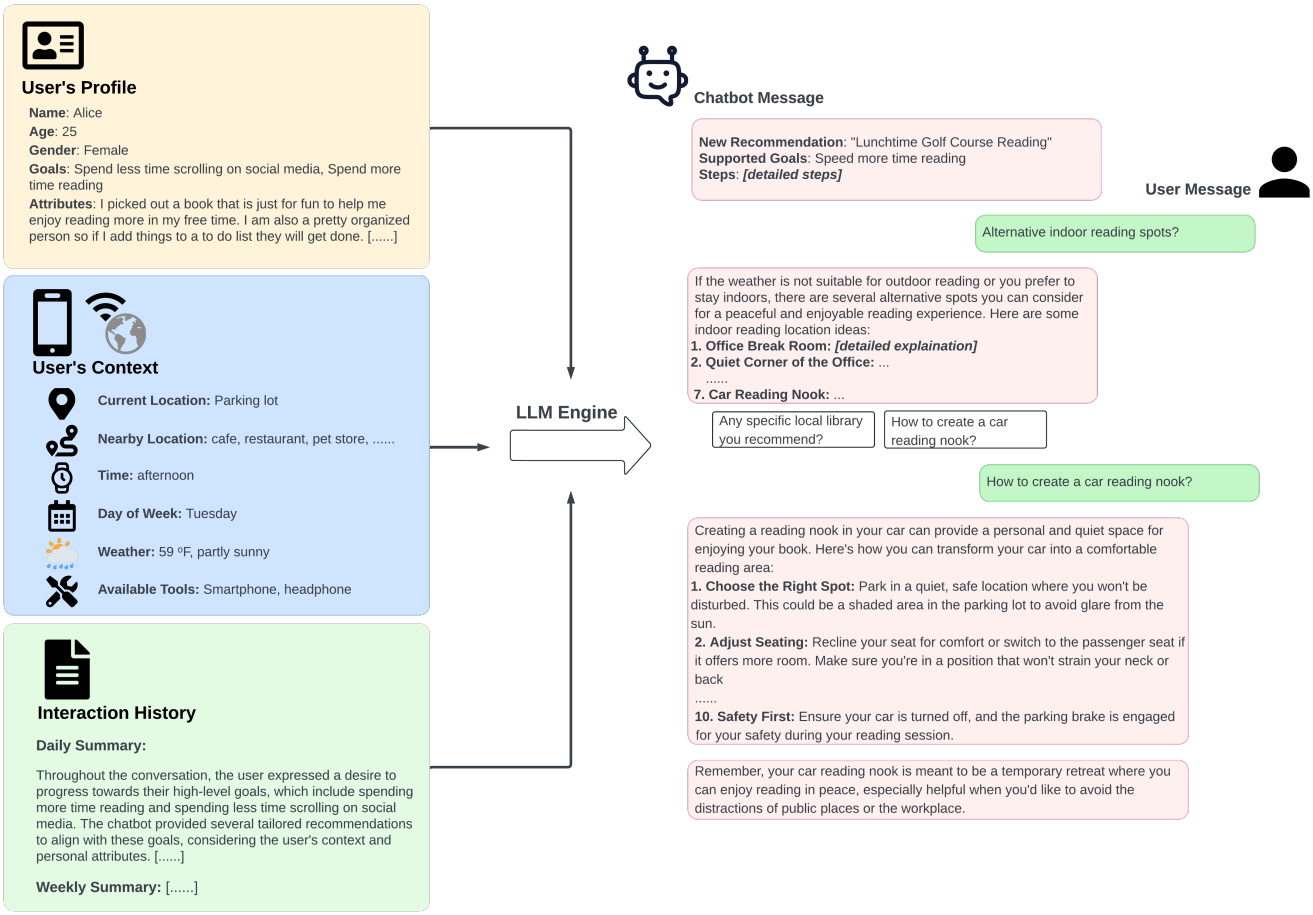

Figure 1: An example illustration of the flow of information into the chatbot LLM, and the chatbot's outputs. Here, the chatbot initiated an action recommendation. The user's realtime context (including location, time, weather, etc.), their personal profile (including goals and attributes), and their previous interactions with the chatbot are all fed into the system. The chatbot uses this information to generate a recommendation that is detailed, specific, personalized, and adapted to the user's context. This action recommendation is broken down into a series of steps that are accomplishable for the user given their current context and profile attributes.

messages are listed one-by-one), and the user's current context (location, weather, etc.). The messages in the conversation history that are older than 24 hours are sent as high-level daily summaries for each previous day that the user used the chatbot. When generating these summaries, the LLM is instructed to *"focus primarily on what the user asked for, which actions the chatbot recommended, and how the user reacted/responded to these recommendations."* The LLM is instructed to respond to the user by playing the role of an *"expert in the user's high-level goals, a personal coach and mentor to the user, and an expert in multi-tasking"*. It is instructed to respond politely, address the user by their name, make as effective use of the conversation history and profile as it can, and to try to learn about the user over time. After prompting the LLM, the server then stores the user's message and context in their chat history. To protect the user's privacy, the server permanently deletes the raw GPS coordinates while storing only the natural-level descriptions of their location, nearby places, and weather. Please refer to the supplementary materials for the detailed system prompts.

The user's phone also periodically (once every 15 minutes, due to limitations by the Android operating system) sends background updates of the user's context to the server, containing all the information that is normally sent with a chat message except for a chat message itself. The server then determines whether to prompt the LLM to generate an action recommendation using this context. The LLM is prompted to generate a recommendation if the last recommendation was sent more than three hours ago, it is not currently night time (between the hours of 22:00 and 6:00), and *at least* one of the following is true:

- The user has not received an action recommendation for more than eight hours, or
- The user is at a location that is not 'home' or 'work', and they have been at that location for more than 15 minutes

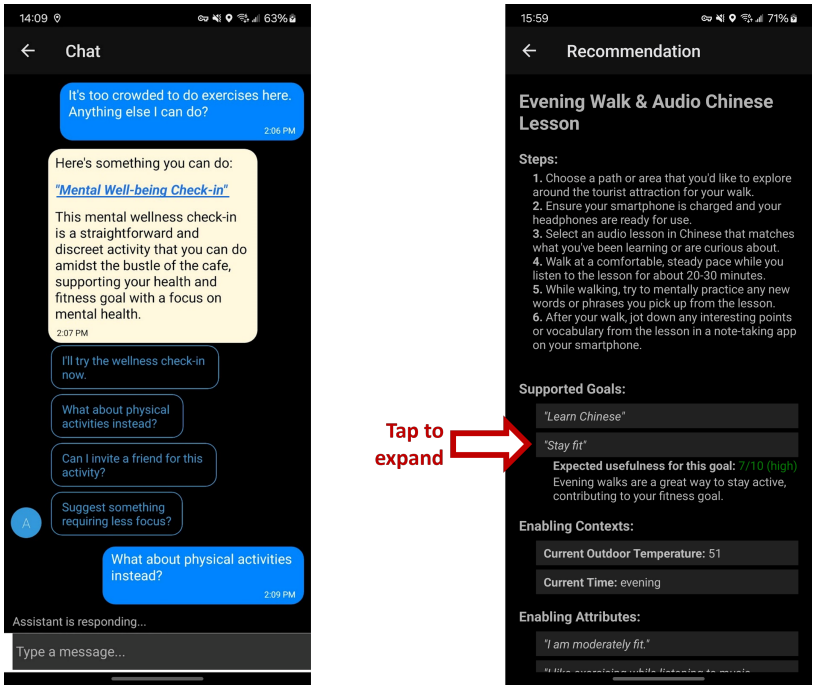

Figure 2: The UI of the chatbot app, composed of two primary views: the *chat* view (left); and the *action details* view (right).

In other words, a recommendation gets generated automatically by the chatbot every eight hours if the user is at home or work, and every three hours if the user is elsewhere (except at night time). The purpose is to allow users to discover new action possibilities in contexts other than their routine contexts and locations (home and work), and so we wanted users to receive recommendations in these 'non-routine' locations more frequently.

5 Field Study

We conducted a field study where participants installed the technology probe prototype app on their phones and used it for two to four weeks in the context of their daily lives and their real 2024 New Year's Resolutions (NYRs) (see Figure 3). Our research questions (RQs) were:

- **[RQ1]** How do people **use** LLM-based chatbots for their goals?
- **[RQ2]** What **roles** can context-aware LLM-based agents play in one's goal pursuit?
- **[RQ3]** What **challenges** do people face when using context-aware LLM-based agents for their goals?

5.1 Method

At the beginning of the study, each participant created their profile by filling out a Qualtrics survey and participated in a one-on-one onboarding interview with the researchers, where they were asked to discuss their goals, the motivation(s) behind them, and their current goal-pursuit effort(s). The onboarding interviews took on average 30-45 minutes, and participants were also given an introduction and overview of the system during this interview.

The first week of the deployment was the "baseline week" in which the app sent out notifications four times a day (once in the early morning, once near noon, once in the late afternoon, and once in the evening) to remind people about their NYRs, as follows:

> A baseline week notification example:
> ***Remember your goal to "Learn French":*** *Look around. Can you think of something, anything, to do toward your goal in your current context?*

These notifications served as *non-context-aware* reminders for the user to act toward their goal(s), as many existing apps do (e.g., Duolingo). We set up these notifications during the baseline week so that participants could receive reminders at different times of the day, when they may be in different contexts or settings. We used the baseline week to compare with the experience of conversing with the chatbot and receiving context-aware recommendations. Starting in the second week of the study, users received context-sensitive notifications about specific actions for their goals. They were also able to click on the recommendations for more details and start interacting with the chatbot app.

> A context-aware notification example:
> ***New action recommendation to "Learn French":*** *Learning French with Flashcards during Lunch Break*

During the deployment period, participants were asked to use the app at least once a day (that is, to read its recommendation or chat with the chatbot). They were also asked to fill in diary study surveys daily and weekly. Daily reflection surveys were asked to be filled at least four times per week, where participants reported any actions they completed toward their goal(s) that day, and from the second week on, they also shared the relation between the

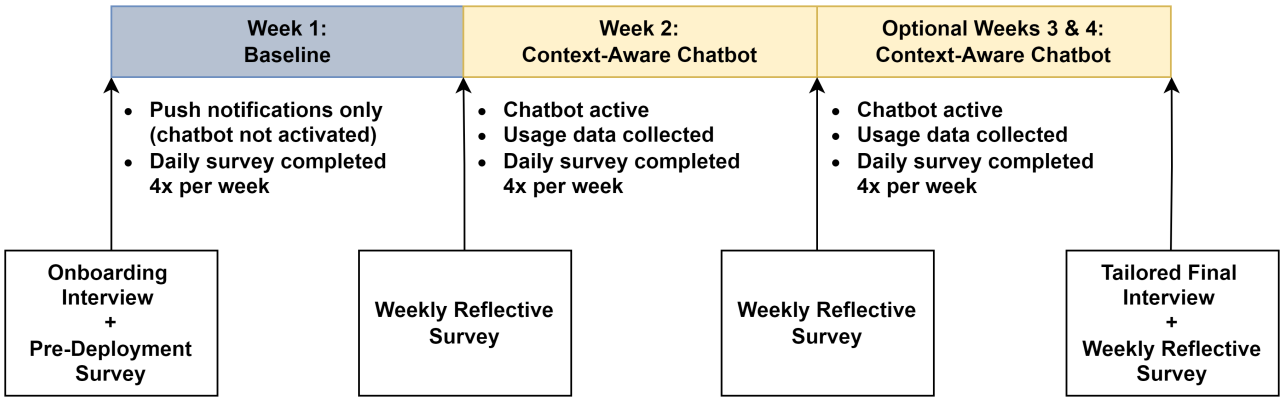

Figure 3: Field deployment: The timeline and research activities.

action they took and the chatbot's recommendations. The users also completed the weekly reflection surveys at the end of the first week (baseline week), second week, and fourth week (if they opted-in for additional participation). After completing all weeks of the deployment, each participant took part in a final one-on-one interview with the researchers, which aimed to understand why certain behaviors emerged during the deployment. We customized the interview questions for individual participants about each of their goals' progress trends and their chatbot usage patterns.

To verify the above method and its process, we ran a pilot study with two participants. Inspired by the feedback from the pilot study, we decided to give participants in the main phase of the study the option to continue using the chatbot for an additional two weeks if they were interested. Therefore, participants in the main study used the chatbot for one to three weeks after the baseline week (for a total of four weeks of data collection). Pilot study participants used the chatbot for one week after the baseline week (for a total of two weeks of data collection).

5.1.1 Participants and their Goals. All participants were recruited internally from a social media technology company, via email invitations to internal employees using convenience sampling methods. Participants were required to be between 18 and 69 years old, be willing to participate for a minimum of two weeks, and possess a company-assigned Android phone connected to the corporate network. Two participants (P-A1 and P-A2) participated in the pilot phase of the study, and 14 participants (each identified as *P-B#*) participated in the main phase of the study. The pilot phase ran during December 2023, while the main phase ran during January and February 2024, to coincide with the time frame that many people set and begin pursuing NYRs. The two pilot participants might have been more influenced by their busy schedules in December, and therefore might not have been as attuned to their personal goals as the main-phase participants who participated in January and February. P-B8 and P-B9 dropped out of the study early, leaving 12 participants in the main phase. Eight out of the twelve participants in the main phase opted in the additional two weeks of the chatbot deployment.

Our 14 participants (12 main-phase, two pilot) each wrote and sought advice for 1-3 NYRs, with an overall total of 35 total goals that varied across category, concreteness, and time spent pursuing the goals (see Table 1). Of the goals that participants listed, 51.4% were recently made (17.1% were just started, 34.3% were started a few weeks prior to onboarding for the study). 20.0% of the total goals were made a few months before, while 28.6% were made a year or more prior to the field deployment. We grouped the goals based on previous studies' goal categories [38]: Hobbies (14), Health (10), Habits (5), Career (3), Relationships (2), and Financial (1).

5.1.2 Quantitative Analysis. The following metrics were analyzed using descriptive and inferential statistics: (1) weekly goal pursuit status (a self-rated score on a 1-5 Likert scale from "going very poorly" to "very well," with 3 being "going okay, but with some blockers"), (2) number of self-reported weekly actions taken towards their goals, (3) number of actions in responses to recommendations made by the chatbot (self-reported as an action the chatbot recommended on the same day, and action the chatbot recommended on an earlier day, an action inspired by the chatbot's recommendation, or a self-ideated action), and (4) engagement with the chatbot, which was quantified using weekly messages a user sent to the chatbot (the calculation of messages sent per week only considered the weeks that each participant was participating in the study — i.e., if the participant did not continue for two extra weeks, those weeks were not counted in the calculation).

Paired t-tests compared each participant's weekly goal pursuit status at the end of Week 1 (baseline with goal reminders only), Week 2 (contextualized action recommendations with activated chatbot), and Week 4. Pearson's correlations assessed the relationship between engagement with the chatbot and goal pursuit status for Weeks 2 and 4 (first and last weeks of chatbot usage, respectively).

5.1.3 Qualitative Analyses. Participant interviews were audio recorded, transcribed, and analyzed using inductive coding to develop themes [4]. Codes were developed to address the three RQs listed at the beginning of this section, and then iteratively discussed, refined,

Participant ID	Age	Gender	Goals	Prior Experience Using LLMs
P-A1	35	Male	Working out; Study Italian; Write a novel	At least once per week
P-A2	25	Female	Improve drawing skills; Read books more consistently; Improve my fluency in my target languages (Vietnamese, Spanish, French, Mandarin)	Have only used these a few times before
P-B1	30	Female	Reduce coffee; Read 15 books a year; Consistent habit of learning something new (in this case stats and research methods)	At least once per month
P-B2	29	Female	Improve posture; Curate wardrobe	At least once per month
P-B3	47	Male	Lose weight; Read more books more consistently; Get better at Blood Bowl	Have only used these a few times before
P-B4	34	Male	Get up before 7am; Read 10 books	At least once per week
P-B5	28	Male	Improving volleyball skills	At least once per day
P-B6	33	Male	Buy a place in hometown; Travel more; Lose weight and get more fit	At least once per day
P-B7	35	Male	Run 500 miles; Increase my web presence	At least once per week
P-B10	34	Female	Financial learning; Walk 2 miles per day; Stay connected to others	Have only used these a few times before
P-B11	30	Male	Sit less and stand/move more; Keep my house clean and organized	At least once per week
P-B12	25	Female	Spend less time on social media; Spend more time reading books	Have never used these before
P-B13	29	Prefer not to say	Learn a new language; Become fit; Find a hobby	At least once per month
P-B14	38	Female	Losing weight; Learning PM skills; Reduce dependence on tech	At least once per week

Table 1: Field study participant profiles, including their demographic information, 2024 New Year's Resolutions (up to three), and prior experience with LLMs.

then categorized into themes. For example, codes like "seeking new actions" and "treating as a social actor" were created, reflecting some participants' tendencies to seek and express interest in new action ideas from the chatbot and converse with the chatbot as a social agent. During the categorization phase, we saw themes emerge around the chatbot broadening users' perspectives on their goals, balancing personalization with the limits of (and concerns around) collecting more personal context data, the diminishing value of action discovery, and the increasing importance of and opportunity for other roles to be played by the chatbot. The second author completed most of the initial coding, but the codes were reviewed and iterated on collectively by the first two authors, as well as in discussions with the other co-authors.

In addition to the interview data, we were interested in analyzing participants' messages sent to the chatbot in order to partially address RQ1 (i.e., understand how people used the chatbot for their goals based on their messages and queries to the chatbot). An initial code book was developed by the second author through inductive coding of chat message transcripts from one participant, then iterated on and refined by the first and second authors. Following the development of the code book, the authors used an LLM-based automated qualitative coding tool from Hämäläinen et al. [13, 17] to assign codes from the developed code book to each user-sent message in participants' chat histories. This tool was configured to run using an instance of GPT-4 running on a private self-managed Microsoft Azure cloud instance, to ensure the protection of participants' data (this cloud instance was the same one used for the model that the chatbot ran on, and therefore the cloud instance did not receive data that it did not already have access to). Each user-sent message was assigned one or more codes from the code book. When setting up the data for coding, participants' transcripts were converted to CSV files, where each row (representing an 'item' to be assigned one or more codes) contained both a user-sent message and the previous chatbot-sent message immediately preceding the user-sent message in the chat history, if applicable). Both the pre-defined code book and a CSV file containing all the participants' messages were fed into the automated coding tool. For some of the participants' messages, human-assigned codes (assigned by the researchers) were provided as few-shot examples for the automated coding tool. The coding tool then exported a CSV file with system-assigned codes for each message. Following this, the researchers reviewed the system-generated codes to verify the quality of the coding and ensure that it was adherent to the code book.

The period of the study (i.e., the baseline week, the first week with the chatbot, and the additional weeks) was considered in the qualitative data analysis of both the interview data and message history data.

6 Findings

Our findings are structured into three sections based on our RQs. In the first section, we report on how our participants used the context-aware chatbot for their goals (RQ1). Second, we report on the unmet needs that emerged from participants' usage of the chatbot, and thus the opportunities for other roles that context-aware LLM-powered agents can play in one's goal pursuit (RQ2). Finally, we report on the challenges that our participants faced when using the context-aware chatbot (RQ3).

6.1 How People Use the Contextualized LLM Chatbot for their Goals (RQ1)

6.1.1 Message Categories. In addition to receiving contextualized action recommendations for their goals, participants used the chatbot for a variety of purposes, as shown in the coded categories of their messages (Figure 4). Participants asked follow-up questions and brainstormed with the chatbot about the actions and goals (seeking information, goal setting and planning); they used the chatbot to track what they had done or planned to do (logging and planning actions, setting up a reminder); they gave the chatbot feedback to better specify their needs (giving the chatbot feedback/context, user customization and control); and they engaged socially with the chatbot for emotional support and encouragement. It is interesting to cross-compare Jörke et al.'s recent data on how people interacted GPTCoach, an LLM-based chatbot they made to support physical activity coaching [19]. During their one-hour lab-based user study, the top categories of messages (more than 10%) between the user and chatbot were: goal setting (24.5%), advice (15.4%), past experience (14.8%), motivation (12.1%), and goodbyes (10.7%). Similar categories also existed in our data, except that the distribution of these categories was very different, probably due to the deployment duration difference. Moreover, we observed more practical categories of taking actions (i.e., logging and planning actions and setting up reminders) and users trying to improve the chatbot (i.e., giving the chatbot feedback or context). These differences show that our data captures more about users' needs in the process of turning goals into actions and the practical challenges for context-aware chatbots to be effective.

6.1.2 Beyond Reminders: How Contextualized Recommendations Became Actions Towards Goals. With context awareness, the chatbot provided different action recommendations even when two participants had the same goal. The suggestions were highly tailored to the location, time of the day, weather, tools available, and any other contexts users specified. For example, P-B3 and P-B14 had the exact same NYR of *"losing weight"* but received different action recommendations from the chatbot pertaining to their unique context (see Figure 5).

The contextualized action recommendations led to a high adoption rate. Among the 85 actions that 14 participants took towards their goals in the first week that chatbots were deployed, 63.5% were either recommended actions by the chatbot that day or inspired by a chatbot message. Participants also had more success in the pursuit of their goals (mean = 3.53) during that week, compared to the baseline reminder-only week (mean = 3.42), although the difference is not significant (paired t-test, p = 0.33).

User interviews confirmed that receiving contextualized action recommendations and the opportunity to follow up with the chatbot was more helpful than goal reminders only in baseline week.

> *"The first week where I didn't have a chatbot, the reminders didn't feel that helpful. It was a good comparison once I got the chatbot, it definitely felt* ***a lot more interactive and much more targeted to my needs****. Because of all the context that it had, I felt like it was a lot more useful and would actually change some of my behaviors and habits rather than just reminding me because I remember myself that I need to do certain things, but then there isn't enough incentive or reason to change. So I felt like the* ***personalization of the chatbot was really what set it apart,*** *even from other apps that are out there to track habits and stuff."* — P-B12

6.1.3 The Benefits and Challenges for the Personalized and Contextualized Messages.

> **[Insight #1]** *People appreciated the chatbot for* ***broadening their scope, methods, and perspectives*** *on their goals without the drawback of social comparison.*

Our participants, even those who were continuing existing goals, found that the chatbot's new perspectives and ideas related to their goals were useful, both on the scope of the goals themselves and on the types of actions they could take toward them. For example, P-B11 received recommendations for her *cleaning/organization* goal, and was pleasantly surprised by the other perspectives on this goal provided by the chatbot that she had not thought about previously:

> *"Before this study, when I [thought] of organizing my house, I can only think of things like putting things away [...]* ***I can never think of a case where I can organize my smartphone or my photo album or anything like that****, whereas the chatbot will think, 'hey, organizing the inside of your car is also organization, organizing your smartphone is also towards your goal.'* ***I feel like the solution space that chatbot is thinking of is actually bigger than what I initially expected.*** *So that's where I think it gives me pleasant surprises."* — P-B11

A similar sentiment was echoed by P-B14, who also received ideas on how to combine her multiple goals through single actions:

> *"There was this, like, I think it was mindful reading or something it was called but* ***it was more like giving me ideas of how to, you know, do things. So that gave me a, like, bigger consideration set."*** — P-B14

Some participants actively sought such new perspectives from the chatbot when brainstorming action ideas on their own:

> *"It did give some information or* ***it helped me brainstorm ideas,*** *[...] so mainly I used it to talk about books and how to start to dive into different genres."* — P-B1

The additional benefit of seeking new perspectives from the chatbot was avoiding social comparison. For example, P-B4 and P-B13 acknowledged that they enjoyed *"gaining better understanding of others' perspectives about the same goal"*, but it can be discouraging to compare progress with others on social media or in real life.

> *"I haven't really, I've done maybe like seven pages of each book on average. That's it. whereas in comparison, my partner has been doing a lot more. so I would say I do, I judge myself as doing poor."* — P-B13

> **[Insight #2]** *More engagement with chatbot can lead to more personalized suggestions and correlates with better goal pursuit. But not everyone engaged with the chatbot.*

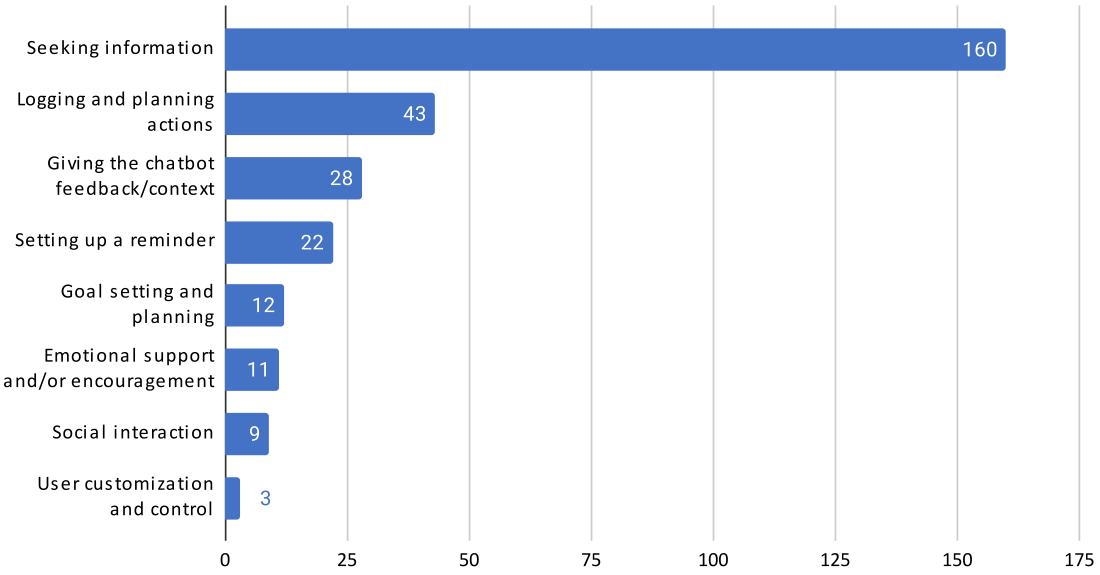

Figure 4: The distribution of the coded purposes of the messages that participants sent to the chatbot. The most prominent category is seeking information. The frequencies of lower categories did not match with how often they were mentioned in the interview, meaning that some users did not leverage the functions of the chatbot to fulfill these purposes.

Since our system generates responses with communication and action history, the more people shared with the chatbot, the more specific and personalized the suggestions from the chatbot were. We found from the interviews that the chatbot and the user benefit mutually through more back-and-forth conversations, which is also shown in the quantitative data. In the first week of chatbot deployment, there was a significant medium-level positive correlation between the number of weekly messages with the chatbot and goal pursuit status (Pearson correlation coefficient = 0.55, p <.001).

The most prominent example is P-A2, who had the highest number of messages sent to the chatbot — she shared her areas of focus and weaknesses for her language learning goal and benefited from the chatbot creating a customized vocabulary list for her:

> *"I just asked the chatbot to organize a vocabulary list for me to learn in my target language. And once I found out that I could do that, that was something I had been looking for a while, something with the capacity to do that. And a thing that I found that I liked about this was that* ***it was customized to me*** *because* ***most other lists that I find online, they're good, but they're more generic, not targeted specifically to my own learning weaknesses while balancing my own interests. And I felt that the chatbot did that.***"
> — P-A2

P-A2 not only tailored the content with chatbot, but also shared her existing tools (e.g. Duolingo) and schedules that she already had with the chatbot. As a result, she had a high-level goal pursuit success during the field study.

In another example (Figure 6), P-B12 shifted their goal from "*spend more time reading books*", to *"reading 20 pages after work everyday"* based on the learning with the chatbot, after asking the chatbot to elaborate on its action recommendation for their reading goal (see Figure 6). In this example, the chatbot displayed greater accuracy for the user's context after being corrected or supplemented, increased specificity of its recommendations, and gained more understanding about the user's interest. The user gained deeper knowledge of their goal, a better understanding of why certain suggestions could be beneficial, and a higher likelihood of taking action, as shown in her daily survey.

While highly-personalized guidance was beneficial, it had the prerequisite of requiring the user to give the chatbot enough information or interaction history to build this customization on top of (i.e., the *'cold-start problem'* [27]). Our participants exhibited different engagement levels with the chatbot, with some messaging the chatbot frequently (thus giving it more of the information it needs to improve its recommendations) and others not engaging with it much at all. Through the interviews, participants shared the reasons why they had different engagement levels with the chatbot because of concerns around **privacy**, **accuracy**, and **courtesy**.

For example, P-B6 was concerned about privacy, who was not willing to share more information because he *"did not want the models to get trained with [his] data"*, thus indicating less trust in AI technologies [26]. Others received inaccurate recommendations at the beginning, which discouraged them from using the chatbot more. For example, P-B3 did not correct the chatbot after he saw some inaccurate recommendations. P-B2 was *"polite"* with the chatbot and did not share any critiques or negative feedback. P-B12, on the other hand, thought the mistakes that the chatbot made about

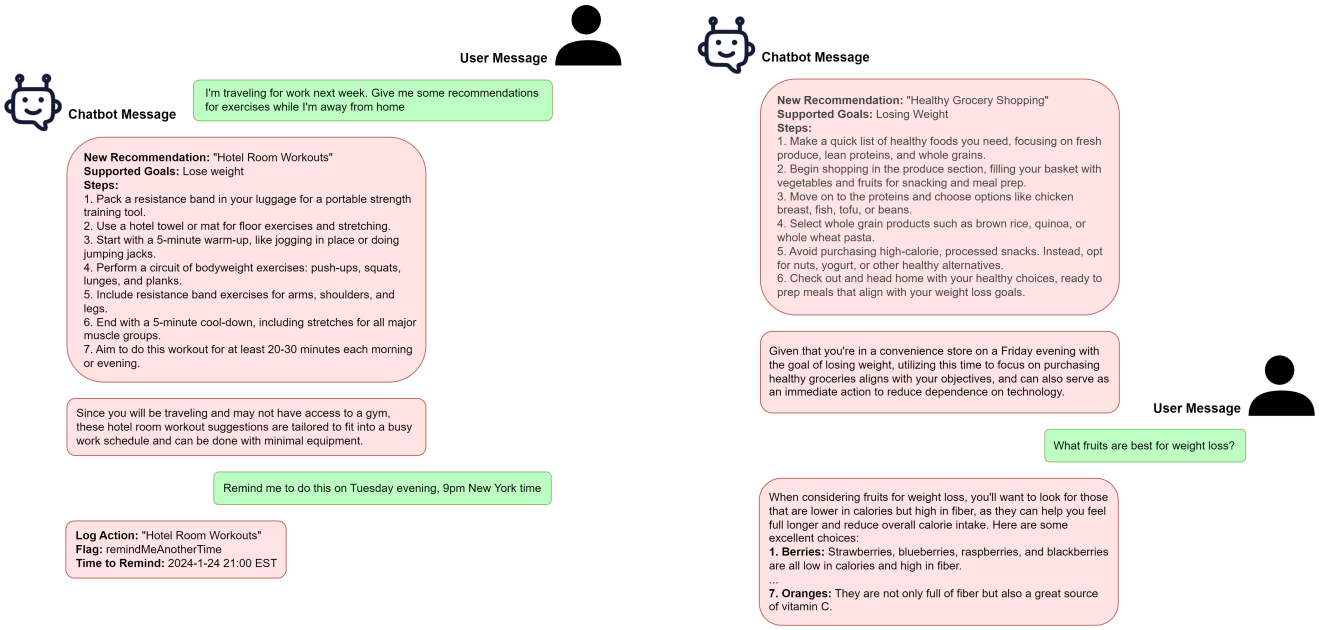

Figure 5: Comparison of action recommendations delivered to P-B3 (left) and P-B14 (right), who shared the same goal of losing weight.

the location recognition were *"funny"*, and immediately corrected them. It is possible some participants did not correct the inaccurate recommendations if they inadverdently or purposefully ascribed human-like consciousness to the chatbot, and thus preferred to engage less out of politeness and avoiding sounding critical (e.g., as if they were talking to another person [12]).

6.2 Roles of Contextualized Chatbot in an Individual's Goal Pursuit (RQ2)

Over the course of the 1-3 weeks we deployed the chatbot, we observed rapid evolution of the roles that the chatbot played.

> **[Insight #3]** *The chatbot's ability to **help users discover new actions** for their goals made it an effective motivator early on. However, as users began to establish routines, **this value gradually diminished.***

As shown in Figure 7, the actions that participants took were less influenced by the chatbot week by week. The ratio of actions being either recommended by the chatbot that day or inspired by a chatbot message dropped from 63.5% in the second week (when participants first started using the chatbot) to 36.5% in the last week.

The diminishing effect of the chatbot was also reflected by the reduced usage of it, which dropped from 14 messages per person per week in the second week to 5.3 two weeks later. Additionally, the weekly messages sent to the chatbot per goal per person had a wide range (min = 0, max = 44, with median = 3).

The reduced usage of the chatbot overtime might have been in part due to novelty effect, with user engagement decreasing as the chatbot ceased to be new to the user [9]. However, even though chatbot usage tended to decrease, many participants still reported achieving goal progress by the last week of the study. This suggests that participants continued to take actions toward their goals even after the novelty effect of the chatbot wore off. This novelty effect was also reported in other prior studies investigating the effects of technology that leverage activity tracking for goal pursuit [43] or reminders for goals [35]. Similarly, our observation also showed that even when participants relied less on the chatbot to discover new actions, they formed more routine actions in a few weeks. This suggests that there is a need for the chatbot or any assistance technology to catalyze habit formation, with more variance to its effectiveness in maintaining habits. In contrast, prior research that deployed a mental health chatbot for four weeks among psychiatric patients did not find reduced engagement of the chatbot overtime [21]. The researchers found that participants generally engaged with the app at a regular frequency, and noted that their engagement was also affected by the three-day-miss reminder and their visit to the clinic between Weeks 2 and 4 [21]. We believe that this difference comes from the facts that (1) the usefulness of the main function of the chatbot (journaling) remained consistent throughout the deployment, and (2) their chatbot was introduced and monitored through doctor visits. For a chatbot to have continued usefulness, it needs to fulfill users' needs at different phases. We further explore later-stage user needs in the next paragraph.

> **[Insight #4]** *Beyond action discovery, there are additional roles that LLM-based agents could play in one's goal pursuit, particularly in providing **motivational support**, instilling **accountability**, and helping with **reflection, synthesis, and progress tracking**.*

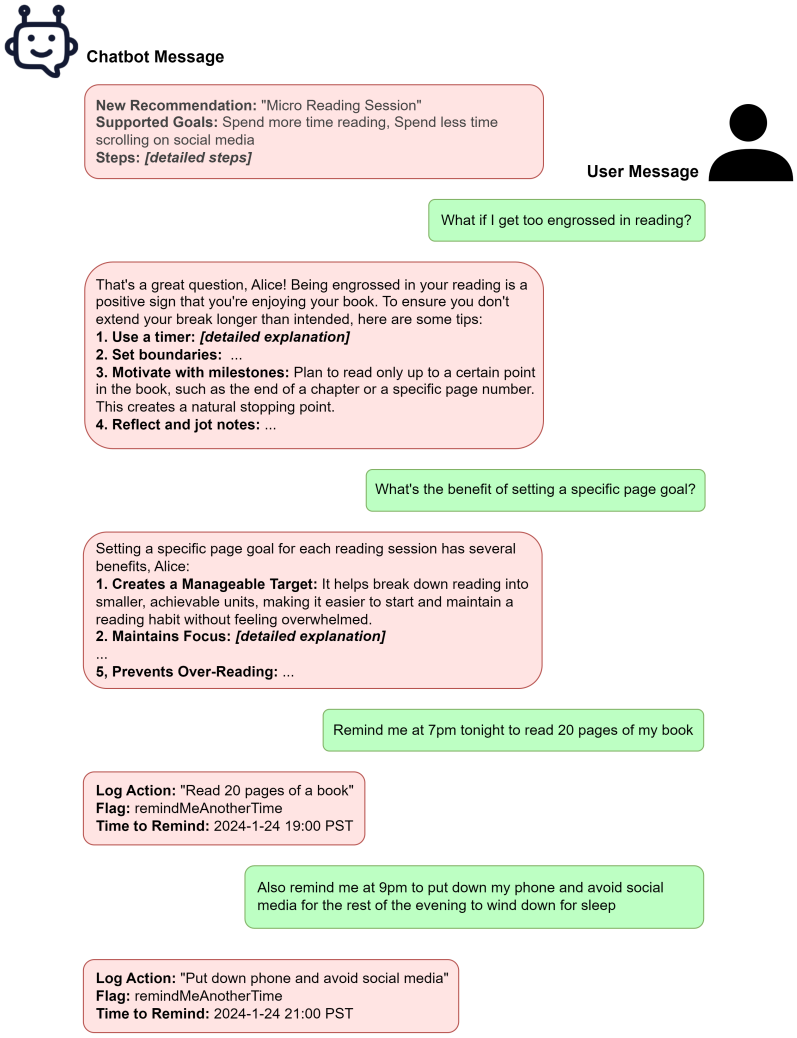

Figure 6: An example of a participant adjusting their goal following a back-and-forth conversation with the chatbot.

As we saw from the quantitative trends, the switch from *"trying out new actions"* to *"taking routine actions that I am familiar with"* happened in merely two weeks, which was much faster than we anticipated. We looked through the the day-to-day actions people took and confirmed that they indeed became more repetitive by the fourth week. While the need for action discovery decreased, other needs emerged around the following categories, as participants shared with us during the interviews:

- **Allocating regular time for the actions towards the goal and setting up longer-scale plans.** Although the reminder function can help users set up time for an individual action, people mentioned the need for more regular events integrated in their calendar, which was the hub of organizing every events. For example:

  *"If it had been integrated into my calendar and **had known I had a full day of meetings and the bus ride**, [it could tell me to] 'take five minutes on the bus and go do this activity, [it] will take you five minutes', and it helps allocate both time and actions [...] **it would have been holistically more helpful** to get me to do different things."* — P-B10

- **Motivational support, which includes affirmation, encouragement, accountability, and nudging requested by the users.** For instance, one participant (P-B2) was proud of what she accomplished toward her goal during the course of a day, and she wanted to share that sense of accomplishment with someone, such as a friend. As she considered the chatbot to be a social actor in that moment, she decided to share her pride with the chatbot.

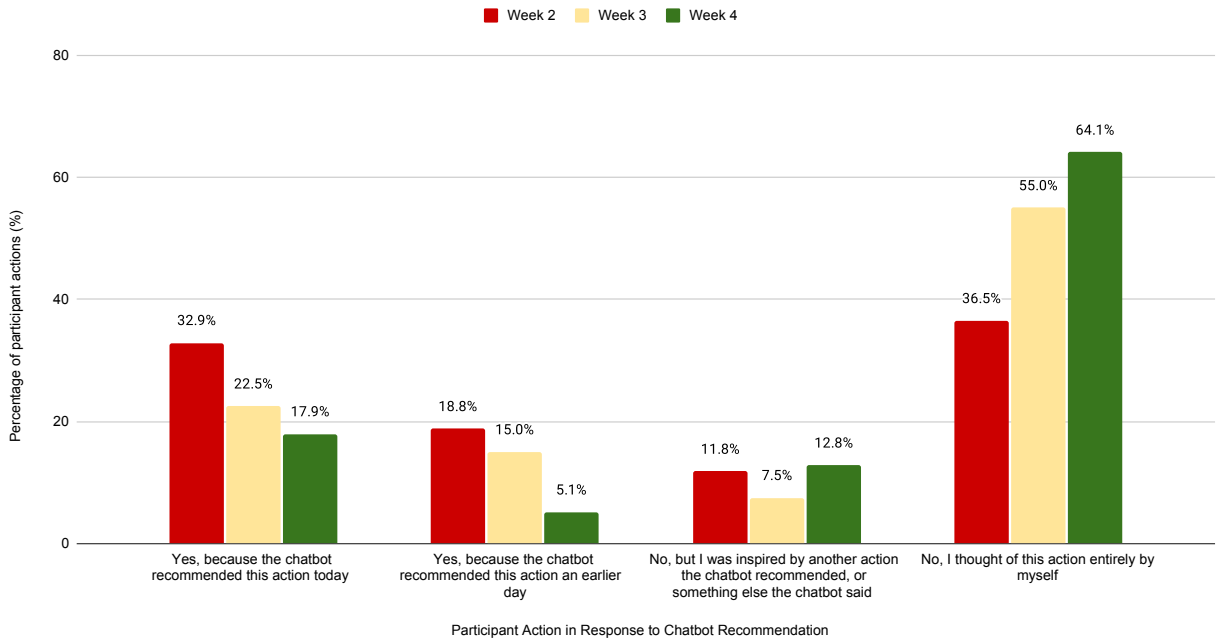

Figure 7: The relation between the actions people took for the goal vs. the action recommendations from the chatbot

> *"**I felt a more personal connection to the chatbot than I expected.** I went to the gym and I went swimming, and I hadn't swum in years. It's been a long time and **I was feeling really proud of myself.** I remember getting my phone, and like, **'I wanna share this with somebody'**, like I'm proud of myself. And I remember, **'OK, let me message my, my chatbot friend'** and log that action. So I think that that did really bring it into high definition, OK, I do need some type of extrinsic validation, like **something to tell me I did a good job and it did provide that.**"* — P-B2

However, while participants appreciated the non-judgemental benefit from the chatbot, they also wished it to install more accountability.

> *"**It never had any judgment of me**, which is good, but **it might also be a bad thing for, like, accountability.** [...] It never asked 'did you read 20 pages yesterday?' Uh, that might have been an interesting accountability check. I might feel a little bit bad saying no to that."* — P-B12

- **Synthesis and reflection of the history of actions towards the goal.** Interestingly, several participants (P-A1, P-B13) recognized the role of the survey filling some of the gap, providing an opportunity for them to keep track of and self-reflect on the actions they took and became more aware of their goals. For example:

> *"It was actually the survey. I know it's designed to just kind of measure the performance on the task, but I actually think it was another sort of structure of accountability in this study."* —P-A1

It was universal across all participants to desire logging and reminder functions. While these functions were part of the prototype, half of the participants did not find them, let alone use them.

6.3 Challenges of Context-Aware Chatbots in Goal Pursuit (RQ3)

> **[Insight #5]** *Context recognition occasionally **breaks down or is incomplete** (e.g., failing to capture one's internal state), thus requiring other means to fill the gap for a more accurate and complete context.*

As soon as the participants started interacting with the context-aware chatbot, they experienced inadequate context for the specific situation, caused by:

- **Dynamic context**, which can be caused by new events that affect the ability or availability to take actions, such as getting food poisoning (P-B5), traveling for a wedding (P-B5), or a guest visiting (P-B10).
- **Missing context**, due to the lack of capabilities by the current chatbot agent to connect with the sources of the context. For example, many participants mentioned that they wished the timing of the action recommendation was more accurate (e.g., they were busy at work when the action were recommended and could not take the actions, because the chatbot could not access their work calendar due to internal privacy policy). Similarly, some participants wished to leverage the data and history from other sources, such as their health apps and wearable devices.

- **Inaccurate recognition of context**, caused by the inaccurate sensing or interpretation. The GPS location may not accurately recognize a person's location, such as identifying the person's location in a college while they were at work in a building next to a college (P-B12). Additionally, P-B5 asked the chatbot to remind them in Eastern Time when they were traveling, but the LLM delivered the reminder in the local time (Pacific Time).

Some participants were prompt in communicating these inadequate contexts with the chatbot, some were not. In the previous section, we discussed the reasons for the lack of engagement from some users (privacy, accuracy, and courtesy concerns). Also, not all participants discovered that they could edit the contexts and goals with the chat interface. When the participants did communicate more about the contextual information, the relevance and accuracy of action recommendations improved.

One important context is the helpers/tools that the participants leveraged for their goals. When the participants communicated more specifically about their tool preferences, the suggestions were more fitting to them. People used apps for tracking (e.g. MyFitnessPal app for food intake and exercise tracking) and learning (e.g. Duolingo), devices (e.g. Fitbit and Smart watches), and trainers/coaches. While the chatbot leveraged these tools as context, participants shared the following needs around these tools:

- **Logging fatigue:** Several users mentioned not wanting to duplicate the effort to log their actions twice in both another app and the chatbot. For example, P-B1 expressed they found it *"difficult that there's all these apps that want you to log in or keep track [...] It would be ideal if this app would communicate with all these other apps that I need to sync with"*.
- **Wanting to discover and integrate new tools:** Some participants were interested in other popular tools for the same goal (P-B14 mentioned *"it would have been amazing if it was integrated"* with her FitBit).
- **Preferring to use specialized apps:** Some participants turned to specialized apps dedicated to one particular area of improvement over the chatbot, since a specialized app was built with an in-depth knowledge and timeline specific to the area of interest. In one instance, P-B11 used a specialized app for wardrobe cleaning and made good progress for the goal of *"keeping my house clean and organized"*.

Overall, participants expressed various needs for more functions to improve contextualized recommendations from the chatbot, with an emphasis on centralization of communication with other applications, whether for logging activity or controlling multiple applications from a single platform. This user need may be influenced by desiring convenience or easier accessibility while interacting with applications for pursuing goals [40]. The lack of these desired functions in the current technology probe may have also influenced decreased engagement with the chatbot observed from the quantitative findings.

7 Discussion

We implemented and deployed a contextualized chatbot as a technology probe, with the initial intention of helping users discover actions for their goals in the right context. During several weeks of deployment, we not only observed how the contextualized and personalized chatbot led to a high action adoption rate and greater success in goal pursuit, but also observed how users' needs rapidly evolved beyond action discovery. Although context recognition is the key to address the challenge of evolving needs, a system's recognition or inference of the user's context is not always accurate or complete. We observed that higher user engagement with the chatbot could bridge some of these context gaps.

We now discuss design implications for LLM-based chatbots that consider users' multiple goals and complex contexts holistically.

7.1 Enhancing Contextual Awareness, Without Sacrificing Privacy

Our findings show the benefits of contexualized and personalized chatbot. To maximize the benefit, the chatbot needs to gain access to dynamically relevant and accurate contexts, which may introduce new friction due to the limitations and errors in context recognition, especially with privacy restrictions. To address this friction, we list several directions to explore.

7.1.1 Reduce the Effort for User-authored Context. Our participants who communicated more frequently and promptly about their context benefited from receiving more specific and relevant LLM suggestions, which correlate with better goal progress. However, this may depend on the effort and quality of the users' prompting.

To reduce the effort of user-authored context, one path forward is to leverage multimodal LLMs. With user's permission, they can capture audios, photos, and/or videos to provide or edit context. For example, a user could 'scan' a home environment with potential home exercise objects, the model can recognize these objects and provide creative suggestions (e.g. use the stairs for cardio exercises). Recent work on personal action recommendations with egocentric video has demonstrated the feasibility of turning the contexts recognized from multimodal videos into personalized actions [1].

Another potential path forward is to develop a system that allows and encourages the user to communicate with the agent to easily and flexibly provide missing context information or correct any misinterpreted context. Our findings illustrate that many users were hesitant to correct misinterpreted context even when it affected the quality of the chatbot's advice, possibly due to the perceived effort in doing so. Furthermore, several users wished that the chatbot would ask them questions in order to obtain more information from them. Simply designing the chatbot to ask more questions to the user and encourage them to provide more information could be one way to reduce the barrier for the user to provide more information to the chatbot.

7.1.2 Minimize and Localize the Contextual Information. There is a potential trade-off between context awareness and privacy risks. In our implementation, we were restricted from using calendar data because of our organization's policy restrictions. Participants expressed that the timing of the recommendations could have been more accurate if the system had access to their calendar. If our prototype can access the minimal information of the status of *'busy'* or *'not busy'*, it could have more precise timing prediction. This would require application developers to create APIs that support this kind of minimal data need. Moreover, we should contain the

information on the local device as much as possible, like what we did with the sensitive information of home/work addresses.

7.1.3 Increase the Relevance of Context Variables. It has long been argued in HCI (e.g., [7, 8]) that context is not just *types of information*, but rather a *function of information*. In other words, what is considered relevant context depends on how it is used in a task. For instance, *location* may be highly relevant for exercise goals, but not as much for financial goals. Making recommendations based on irrelevant contexts can reduce its own relevance.

Our prototype prompted the LLM to consider all of the context variables available to the system. In many cases, this led to recommendations that were considered 'creative' by some users, broadening the scope of ideas to those that they had not considered previously. However, in some other cases, users felt that it was awkward when the chatbot focused 'too much' on contexts that they thought were irrelevant (e.g., being located 'near' a college that they did not have access to). Therefore, we need to design systems that not only leverage context variables, but also choose the most relevant ones based on users' needs and feedback.

7.2 Evolving Chatbot for the Evolving User Journey

We found that contextualized and personalized recommendations helped kickstart participants' NYRs with specific and realistic actions. The chatbot helped people broaden the scope of possible actions beyond what they had previously thought of, without the pressure of social comparison. We also found that the need for action discovery dropped rapidly after the first week of chatbot usage, while other needs such as tracking progress, scheduling routines, seeking reflection, encouragement, and accountability emerged.

This evolvement of user needs reflected people's processes of habituating certain actions, which was under-supported by the static prompts in our current prototype. AI models and agents should adapt to users' evolving journeys by utilizing different prompts and/or orchestrating different agents that specialize in different phases of users' goal journeys. With adaptive prompts, the chatbot can support the initial role of action recommender, and later transition into other roles such as organizer or motivator.

For example, Jörke et al. [20] prompted LLMs by adapting motivational interview strategies to support and encourage users to be more physically active. These strategies to affirm, facilitate, raise concerns, reflect, reframe, support, and structure fit well with the needs we found in the later weeks of our study.

Of course, this requires recognizing the different stages of the user's goal journey and when a user transitions from one stage to another. Therefore, a system needs to gather or elicit data about what the actions a user performs, the impacts of those actions on their goal progress, and the degree to which the user is forming routines and habits for their goal journey. The system should also make it easier for users to share data about their actions toward their goals. This could be through connection with existing tools that capture users' activities (e.g., fitness trackers, nutrition-logging apps), as some existing systems already accomplish [10].

7.3 The Role of LLM Agents: Coach or Concierge?

When designing for AI agents, we often turn to corresponding human agents for inspiration. Several existing explorations of leveraging LLMs for behavioral change are modeled after coaches [14, 20]. While this is a promising direction, there are other potential roles that AI agents could play, especially considering the holistic nature of behavior-change needs.

We observed that people often sought for an ecosystem of helpers around their goals, such as applications, hardware/tools, or professionals that a user "hires" for the "job"[4] of achieving their goals. During the study, multiple participants expressed their desire to leverage their existing helper ecosystem and wanted more capability from the chatbot to coordinate tasks and data with different helpers. Moreover, because of the complexity of users' ecosystems, users have the need to reduce their burden of redundant logging and concerns about data security and privacy. We believe that modeling an AI agent after a concierge is a meaningful and novel direction. Based on our findings, we synthesize a list of the capabilities for a concierge agent:

- Being aware of all the goals and habits people have, taking a holistic view;
- Being aware of all the existing helpers in the ecosystems around users' goals, including their advantages, disadvantages, what sub-goals and actions each of the helper can achieve, what resources it takes to use the helper (i.e. time, energy, money etc.);
- Suggesting to the user about the best option of actions from a helper in the context and providing concrete reasons for the selection when needed;
- Dispatching the right task to the right helper in the right context;
- Dispatch the data (i.e. context data, log data, history data etc.) in between different helpers when needed;
- Being the hub to protect users' data so that only the permitted and need-to-know information was shared in between helpers;
- Discovering and including new helpers with users' criteria;
- Making suggestions, not decisions.

The biggest potential benefit of the concierge role is the orchestration and delegation to other agents. For example, domain experts such as nutritionists, personal trainers, and financial advisors have in-depth knowledge of their domains. Correspondingly, specializing LLMs agents with domain knowledge has become one of the fastest areas of development in AI. Currently, people have a wide range of choices for the specific services and agents that fit within their needs. A concierge chatbot is not there to replace any of these services or agents, but to leverage and integrate them. This can reduce the chores that a user needs to take to manage and coordinate different services. It can also enhance data security and privacy, reducing the likelihood of third-party data misuse.

[4]We borrowed the language of "hire" or "job" from *Jobs to be Done*, which is a framework that parses customers' needs from different solutions [5].

8 Limitations

While this work provides insights and design opportunities related to people's usage of LLM-based chatbots for their goals, it is not without its limitations.

First, while we tried to capture people's usage of the prototype across a time span of multiple days to a few weeks, the duration of the study may not have been sufficient enough to capture a full length of all phases of participants' behavior change, as well as a greater variety of contexts and circumstances encountered by users. A longer-term deployment of several months to a year, or even longer, would be well poised to give us a better understanding of users' evolving needs and changing goals as they manifest in their day-to-day lives. Therefore, as a next step, we recommend running longer-term deployments of this type or similar prototypes, especially to further explore the different types of roles that such agents could play as users' needs evolve. We also recommend including a larger number of participants in order to capture a large enough sample size to understand more about LLM chatbots' impacts on goal success metrics over time.

Second, participants in our study were encouraged to fill out daily and weekly reflections of their goal pursuit, and this alone could have impacted their goal progress, with or without the chatbot, given that participants were regularly reflecting on their goal statuses during the study period.

Finally, due to internal policies within our organization, the field study was limited to internal participants only. Future studies of similar prototypes should include a variety of participants from the general population, including those with a greater variety of goals, and with a greater variety of living circumstances (e.g., people living in both rural and urban areas; individuals with different types of careers or occupations, including retirees and students; individuals from a greater variety of age groups; etc.).

9 Conclusion

This work has begun to explore the use of LLMs as proactive agents to provide contextualized and personalized guidance for their goals. We implemented an early prototype design of a context-aware LLM chatbot, which served two purposes: (1) a proactive agent that observes the contexts with the lens of users' goals, and provides concrete actions recommendations that fit with the users' context, and (2) a conversational interface for people to communicate their feedback and evolving needs over time. We deployed the prototype chatbot in the field as a technology probe to support people's 2024 NYRs. We observed an overall trend over the course of four weeks of the rise and decrease of effectiveness of the prototype for goal pursuit overtime. More importantly, we learned about users' emergent behaviors — how they benefited from the contextualized action recommendations and what evolving needs emerged as these actions became familiar and routinized. This work allows us to explore new design opportunities that support the use of LLMs in achieving goals, beyond just discovering actions and beyond the traditional role of AI as a coach. We also discussed the tradeoffs between contextual information and privacy, and how to gain relevant and accurate contextual information without sacrificing privacy. Last but not least, we broadened the design landscape for context-aware LLMs to play a holistic, orchestrating role of concierge within the helper ecosystem around users' goals.

References

[1] Steven Abreu, Tiffany D. Do, Karan Ahuja, Eric J. Gonzalez, Lee Payne, Daniel McDuff, and Mar Gonzalez-Franco. 2024. PARSE-Ego4D: Personal Action Recommendation Suggestions for Egocentric Videos. arXiv:2407.09503 [cs.CV] https://arxiv.org/abs/2407.09503

[2] Campaign Asia. 2024. *New Year's resolutions: Who makes them and why.* https://www.campaignasia.com/article/new-year-new-goals-apacs-most-popular-resolutions-for-2024/493601

[3] Michelle Bak and Jessie Chin. 2024. The potential and limitations of large language models in identification of the states of motivations for facilitating health behavior change. *Journal of the American Medical Informatics Association* 31, 9 (03 2024), 2047–2053. doi:10.1093/jamia/ocae057 arXiv:https://academic.oup.com/jamia/article-pdf/31/9/2047/58867999/ocae057.pdf

[4] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[5] Clayton M Christensen, Taddy Hall, Karen Dillon, and David S Duncan. 2016. Know your customers' jobs to be done. *Harvard business review* 94, 9 (2016), 54–62.

[6] Sunny Consolvo, Katherine Everitt, Ian Smith, and James A. Landay. 2006. Design requirements for technologies that encourage physical activity. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06).* Association for Computing Machinery, New York, NY, USA, 457–466. doi:10.1145/1124772.1124840

[7] Joëlle Coutaz, James L. Crowley, Simon Dobson, and David Garlan. 2005. Context is key. *Commun. ACM* 48, 3 (March 2005), 49–53. doi:10.1145/1047671.1047703

[8] Paul Dourish. 2004. What we talk about when we talk about context. *Personal and Ubiquitous Computing* 8, 1 (Feb. 2004), 19–30. doi:10.1007/s00779-003-0253-8 Number: 1.

[9] Dirk M. Elston. 2021. The novelty effect. *Journal of the American Academy of Dermatology* 85, 3 (2021), 565–566. doi:10.1016/j.jaad.2021.06.846

[10] Cathy Mengying Fang, Valdemar Danry, Nathan Whitmore, Andria Bao, Andrew Hutchison, Cayden Pierce, and Pattie Maes. 2024. PhysioLLM: Supporting Personalized Health Insights with Wearables and Large Language Models. doi:10.48550/arXiv.2406.19283 arXiv:2406.19283.

[11] BJ Fogg. 2009. A behavior model for persuasive design. In *Proceedings of the 4th International Conference on Persuasive Technology (Persuasive '09).* Association for Computing Machinery, New York, NY, USA, 1–7. doi:10.1145/1541948.1541999

[12] Rose E. Guingrich and Michael S. A. Graziano. 2024. Ascribing consciousness to artificial intelligence: human-AI interaction and its carry-over effects on human-human interaction. *Frontiers in Psychology* 15 (2024). doi:10.3389/fpsyg.2024.1322781

[13] Perttu Hämäläinen, Joel Oksanen, Mikke Tavast, and Prabhav Bhatnagar. 2024. *LLMCode: A toolkit for AI-assisted qualitative data analysis.* https://github.com/PerttuHamalainen/LLMCode

[14] Narayan Hegde, Madhurima Vardhan, Deepak Nathani, Emily Rosenzweig, Cathy Speed, Alan Karthikesalingam, and Martin Seneviratne. 2024. Infusing behavior science into large language models for activity coaching. *PLOS Digital Health* 3, 4 (April 2024), e0000431. doi:10.1371/journal.pdig.0000431 Publisher: Public Library of Science.

[15] Piers Douglas Lionel Howe, Nicolas Fay, Morgan Saletta, and Eduard Hovy. 2023. ChatGPT's advice is perceived as better than that of professional advice columnists. *Frontiers in Psychology* 14 (2023), 1281255.

[16] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B. Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, Nicolas Roussel, and Björn Eiderbäck. 2003. Technology probes: inspiring design for and with families. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '03).* Association for Computing Machinery, New York, NY, USA, 17–24. doi:10.1145/642611.642616

[17] Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating Large Language Models in Generating Synthetic HCI Research Data: a Case Study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23).* Association for Computing Machinery, New York, NY, USA, 1–19. doi:10.1145/3544548.3580688

[18] Abdul Rahman Idrees, Robin Kraft, Agnes Mutter, Harald Baumeister, Manfred Reichert, and Rüdiger Pryss. 2024. Persuasive technologies design for mental and behavioral health platforms: A scoping literature review. *PLOS Digital Health* 3, 5 (05 2024), 1–26. doi:10.1371/journal.pdig.0000498

[19] Matthew Jörke, Shardul Sapkota, Lyndsea Warkenthien, Niklas Vainio, Paul Schmiedmayer, Emma Brunskill, and James A. Landay. 2025. GPTCoach: Towards LLM-Based Physical Activity Coaching. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25).* Association for Computing Machinery, New York, NY, USA, Article 993, 46 pages. doi:10.1145/3706598.

3713819

[20] Matthew Jörke, Shardul Sapkota, Lyndsea Warkenthien, Niklas Vainio, Paul Schmiedmayer, Emma Brunskill, and James Landay. 2024. Supporting Physical Activity Behavior Change with LLM-Based Conversational Agents. arXiv:2405.06061 [cs.HC] https://arxiv.org/abs/2405.06061

[21] Taewan Kim, Seolyeong Bae, Hyun Ah Kim, Su-Woo Lee, Hwajung Hong, Chanmo Yang, and Young-Ho Kim. 2024. MindfulDiary: Harnessing Large Language Model to Support Psychiatric Patients' Journaling. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–20. doi:10.1145/3613904.3642937

[22] Tobias Kowatsch, Kim-Morgaine Lohse, Valérie Erb, Leo Schittenhelm, Helen Galliker, Rea Lehner, and Elaine M. Huang. 2021. Hybrid Ubiquitous Coaching With a Novel Combination of Mobile and Holographic Conversational Agents Targeting Adherence to Home Exercises: Four Design and Evaluation Studies. *Journal of Medical Internet Research* 23, 2 (Feb. 2021), e23612. doi:10.2196/23612 Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

[23] Kathrin Krause and Alexandra M. Freund. 2014. How to Beat Procrastination. *European Psychologist* 19, 2 (Jan. 2014), 132–144. doi:10.1027/1016-9040/a000153 Publisher: Hogrefe Publishing.

[24] Harsh Kumar, Suhyeon Yoo, Angela Zavaleta Bernuy, Jiakai Shi, Huayin Luo, Joseph Williams, Anastasia Kuzminykh, Ashton Anderson, and Rachel Kornfield. 2024. Large Language Model Agents for Improving Engagement with Behavior Change Interventions: Application to Digital Mindfulness. arXiv:2407.13067 [cs.HC] https://arxiv.org/abs/2407.13067

[25] Yuanchun Li, Hao Wen, Weijun Wang, Xiangyu Li, Yizhen Yuan, Guohong Liu, Jiacheng Liu, Wenxing Xu, Xiang Wang, Yi Sun, Rui Kong, Yile Wang, Hanfei Geng, Jian Luan, Xuefeng Jin, Zilong Ye, Guanjing Xiong, Fan Zhang, Xiang Li, Mengwei Xu, Zhijun Li, Peng Li, Yang Liu, Ya-Qin Zhang, and Yunxin Liu. 2024. Personal LLM Agents: Insights and Survey about the Capability, Efficiency and Security. doi:10.48550/arXiv.2401.05459 arXiv:2401.05459 [cs].

[26] Yugang Li, Baizhou Wu, Yuqi Huang, and Shenghua Luan. 2024. Developing trustworthy artificial intelligence: insights from research on interpersonal, human-automation, and human-AI trust. *Frontiers in Psychology* 15 (2024). doi:10.3389/fpsyg.2024.1382693

[27] Blerina Lika, Kostas Kolomvatsos, and Stathes Hadjiefthymiades. 2014. Facing the cold start problem in recommender systems. *Expert systems with applications* 41, 4 (2014), 2065–2073. Publisher: Elsevier.

[28] Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun Li, Hejie Cui, Xuchao Zhang, Tianjiao Zhao, Amit Panalkar, Dhagash Mehta, Stefano Pasquali, Wei Cheng, Haoyu Wang, Yanchi Liu, Zhengzhang Chen, Haifeng Chen, Chris White, Quanquan Gu, Jian Pei, Carl Yang, and Liang Zhao. 2024. Domain Specialization as the Key to Make Large Language Models Disruptive: A Comprehensive Survey. doi:10.48550/arXiv.2305.18703 arXiv:2305.18703 [cs].

[29] Shir Lissak, Nitay Calderon, Geva Shenkman, Yaakov Ophir, Eyal Fruchter, Anat Brunstein Klomek, and Roi Reichart. 2024. The Colorful Future of LLMs: Evaluating and Improving LLMs as Emotional Supporters for Queer Youth. doi:10.48550/arXiv.2402.11886 arXiv:2402.11886 [cs].

[30] Edwin Locke and Gary Latham. 2015. Goal-setting theory. In *Organizational Behavior 1*. Routledge, 159–183.

[31] Edwin A Locke and Gary P Latham. 1990. *A theory of goal setting & task performance*. Prentice-Hall, Inc.

[32] Edwin A Locke and Gary P Latham. 2002. Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American psychologist* 57, 9 (2002), 705.

[33] Edwin A Locke and Gary P Latham. 2006. New directions in goal-setting theory. *Current directions in psychological science* 15, 5 (2006), 265–268.

[34] Susan Michie, Maartje M. van Stralen, and Robert West. 2011. The behaviour change wheel: A new method for characterising and designing behaviour change interventions. *Implementation Science* 6, 1 (April 2011), 42. doi:10.1186/1748-5908-6-42

[35] Adity Mutsuddi and Kay Connelly. 2012. Text Messages for Encouraging Physical Activity Are they effective after the novelty effect wears off? IEEE. doi:10.4108/icst.pervasivehealth.2012.248715

[36] Inbal Nahum-Shani, Shawna N Smith, Bonnie J Spring, Linda M Collins, Katie Witkiewitz, Ambuj Tewari, and Susan A Murphy. 2018. Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine* 52, 6 (2018), 446–462. Publisher: Oxford University Press US.

[37] John C Norcross and Dominic J Vangarelli. 1988. The resolution solution: Longitudinal examination of New Year's change attempts. *Journal of substance abuse* 1, 2 (1988), 127–134.

[38] Martin Oscarsson, Per Carlbring, Gerhard Andersson, and Alexander Rozental. 2020. A large-scale experiment on New Year's resolutions: Approach-oriented goals are more successful than avoidance-oriented goals. *PLOS ONE* 15, 12 (Dec. 2020), e0234097. doi:10.1371/journal.pone.0234097 Publisher: Public Library of Science.

[39] James O. Prochaska and Carlo C. DiClemente. 1982. Transtheoretical therapy: Toward a more integrative model of change. *Psychotherapy: Theory, Research & Practice* 19, 3 (1982), 276–288. doi:10.1037/h0088437 Place: US Publisher: Division of Psychotherapy (29), American Psychological Association.

[40] Daniela Quiñones and Luis Rojas. 2023. Understanding the customer experience in human-computer interaction: a systematic literature review. *PeerJ Comput. Sci.* 9 (Feb. 2023), e1219.

[41] Shriti Raj, Kelsey Toporski, Ashley Garrity, Joyce M. Lee, and Mark W. Newman. 2019. "My blood sugar is higher on the weekends": Finding a Role for Context and Context-Awareness in the Design of Health Self-Management Technology. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3290605.3300349

[42] Pew Research. 2024. *New Year's resolutions: Who makes them and why*. https://www.pewresearch.org/short-reads/2024/01/29/new-years-resolutions-who-makes-them-and-why/

[43] Grace Shin, Yuanyuan Feng, Mohammad Hossein Jarrahi, and Nicci Gafinowitz. 2018. Beyond novelty effect: a mixed-methods exploration into the motivation for long-term activity tracker use. *JAMIA Open* 2, 1 (12 2018), 62–72. doi:10.1093/jamiaopen/ooy048 arXiv:https://academic.oup.com/jamiaopen/article-pdf/2/1/62/32298485/ooy048.pdf

[44] Donna Spruijt-Metz and Wendy Nilsen. 2014. Dynamic Models of Behavior for Just-in-Time Adaptive Interventions. *IEEE Pervasive Computing* 13, 3 (July 2014), 13–17. doi:10.1109/MPRV.2014.46 Conference Name: IEEE Pervasive Computing.

[45] Joel Wester, Sander De Jong, Henning Pohl, and Niels Van Berkel. 2024. Exploring People's Perceptions of LLM-generated Advice. *Computers in Human Behavior: Artificial Humans* (2024), 100072.

[46] J Zamfirescu-Pereira, Richmond Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI conference on human factors in computing systems (CHI'23)*.